Etiquetagem morfossintática multigênero para o português do Brasil segundo o modelo "Universal Dependencies"

Emanuel Huber Silva[1,3,4], **Thiago Alexandre Salgueiro Pardo**[1], **Norton Trevisan Roman**[2]

[1]Núcleo Interinstitucional de Linguística Computacional (NILC),
Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo (USP)

[2]Escola de Artes, Ciências e Humanidades - Universidade de São Paulo (USP)

[3]Centro de Inovação CESAR

[4]Departamento de Engenharia da Computação - Facens

emanuel.huber@usp.br, taspardo@icmc.usp.br, norton@usp.br

Abstract. *Part of speech tagging is a process that seeks to identify the grammatical classes of words and symbols (tokens) in a sentence. For Brazilian Portuguese, there is a variety of approaches using corpora of the journalistic genre with different tagsets. In this paper, we present results better than the current state of the art, investigating tagging methods and evaluating their ability to perform multi-genre analysis in corpora of journalistic, academic and user-generated content genres. To do so, we use the Universal Dependencies model. Finally, we present a qualitative assessment of the systematic tagging errors made in the process.*

Resumo. *A etiquetagem morfossintática é um processo que busca identificar as classes gramaticais de palavras e símbolos (tokens) em uma sentença. Para o português brasileiro, há uma variedade de trabalhos utilizando corpora de gênero jornalístico com diferentes conjuntos de etiquetas. Neste artigo, apresentamos resultados que superam o estado da arte atual, investigando métodos de etiquetagem e avaliando sua capacidade de análise multigênero em corpora dos gêneros jornalístico, acadêmico e de "user-generated content". Para tanto, usamos o modelo "Universal Dependencies". Por fim, apresentamos uma avaliação qualitativa dos erros sistemáticos cometidos pelo modelo.*

1. Introdução

A área de Processamento de Línguas Naturais (PLN) busca automatizar tarefas que envolvam a interpretação e a geração de língua natural [Jurafsky e Martin 2009]. Em várias dessas tarefas, faz-se necessário utilizar características linguísticas dos documentos, como as classes gramaticais de palavras e símbolos (ou etiquetas morfossintáticas dos *tokens* – do inglês, *part of speech tags*) e a estruturação sintática das sentenças.

Apesar da dominância atual das abordagens neurais e dos grandes modelos de língua, que na maioria das vezes processam textos em suas formas originais sem anotação linguística sofisticada, há muitas evidências da importância de informações linguísticas em PLN. Por exemplo, [Lin et al. 2021] combinam etiquetas morfossintáticas com representações vetoriais para aprimorar um analisador de opiniões baseado em aspectos. [Zhao et al. 2019], na frente de sumarização automática, demostram a importância

de utilizar informações lexicais e de etiquetas morfossintáticas em conjunto com mecanismos de atenção. [Cabral et al. 2022], por sua vez, fazem uso desses conhecimentos no desenvolvimento de um sistema de extração de informação aberta para o português. [Garimella et al. 2019], em um estudo socio-linguístico, demonstram que há diferenças gramaticais em textos escritos por homens e mulheres.

Motivadas pela importância desse tipo de conhecimento em PLN, há várias iniciativas clássicas e mais recentes para o desenvolvimento de recursos e ferramentas relacionados para o processamento computacional da língua portuguesa. Pode-se citar, por exemplo, a amplamente conhecida Floresta Sintá(c)tica [Afonso et al. 2002] e o *treebank* Porttinari [Pardo et al. 2021], o léxico de língua geral PortiLexicon-UD [Lopes et al. 2022], o etiquetador morfossintático LX-Tagger [Branco e Silva 2004] e o etiquetador do estado da arte de [Fonseca et al. 2015] treinado com o corpus de referência Mac-Morpho [Aluísio et al. 2003], assim como o conhecido *parser* PALAVRAS [Bick 2000], entre muitas outras pesquisas relevantes.

Visando a contribuir nesta frente e avançar a fronteira do conhecimento, este artigo foca na tarefa de etiquetagem morfossintática para o português, mas trazendo ambições maiores. Por um lado, são investigados métodos variados e do estado da arte para conjuntos de dados de referência em português, avaliando-se a capacidade de análise multigênero dos métodos. Objetiva-se, com isso, o desenvolvimento de um etiquetador de alta acurácia e de amplo uso, possibilitando o desenvolvimento de aplicações de PLN mais robustas. Para tanto, utilizam-se os corpora Porttinari [Pardo et al. 2021], DANTEStocks [Di Felippo et al. 2021] e PetroGold [Souza et al. 2021], dos gêneros jornalístico, gerado por usuário (do inglês, *User-Generated Content* - UGC) e acadêmico (do domínio de óleo e gás), respectivamente. Por outro lado, explora-se o modelo *Universal Dependencies* (UD) [de Marneffe et al. 2021], de ampla aceitação, inclusive para o português [Rademaker et al. 2017]. Mostramos que nossos melhores resultados ultrapassam 99% de acurácia e que é possível produzir um etiquetador morfossintático multigênero de alta acurácia, superando o estado da arte. Mais do que isso, na análise qualitativa realizada, evidencia-se que muitos dos erros remanescentes são linguisticamente plausíveis.

O restante desse trabalho está organizado como segue. Na Seção 2, os trabalhos relacionados são sucintamente apresentados. Na Seção 3, os corpora utilizados são introduzidos. Os experimentos realizados e os resultados atingidos são relatados nas Seções 4 e 5. Por fim, a Seção 6 conclui esse trabalho.

2. Trabalhos relacionados

Há vários trabalhos em etiquetagem morfossintática para o português, dos quais destacamos alguns. [Fonseca et al. 2015] utilizam uma rede neural com representações vetoriais das palavras e atributos linguísticos adicionais (como capitalização e sufixos) para prever suas etiquetas. Os autores utilizam diferentes versões do corpus jornalístico Mac-Morpho [Aluísio et al. 2003], atingindo $97,57\%$ de acurácia (ou seja, a proporção de *tokens* corretamente classificados). Utilizando o mesmo corpus, [de Sousa e Lopes 2019] avaliam as Redes Neurais Recorrentes (RNRs) bidirecionais com representações vetoriais em nível de palavra e caractere. Essa abordagem alcançou $97,36\%$ de acurácia. [Domingues 2011] apresenta um etiquetador que utiliza o aprendizado baseado em transformações para os gêneros jornalístico e acadêmico. Foram utilizados um léxico

Tabela 1. Exemplos dos corpora selecionados

Corpus	Exemplo
Porttinari-base	Foram/AUX avaliados/VERB 5.281/NUM municípios/NOUN ,/PUNCT ou/CCONJ 95/NUM %/SYM de/ADP o/DET total/NOUN de/ADP 5.569/NUM existentes/ADJ em/ADP o/DET Brasil/NOUN ./PUNCT
DANTEStocks	BBAS3/PROPN comprar/VERB por/ADP R$/SYM 20,05/NUM indicado/VERB em/ADP 27/02/2014/NUM 10:41/NUM http://t.co/zJRs3Eeyz9/SYM
PetroGold	Segundo/ADP Luiz/PROPN &/PROPN Silva/PROPN (/PUNCT 1995/NUM)/PUNCT estas/DET feições/NOUN definem/VERB a/DET maioria/NOUN de/ADP os/DET lineamentos/NOUN em/ADP mapas/NOUN magnéticos/ADJ ./PUNCT

para o tratamento de nomes próprios, regras manuais e a saída de outros dois etiquetadores disponíveis na literatura. Além do Mac-Morpho, o trabalho também utilizou o Bosque (que integra a Floresta Sintá(c)tica) para o gênero jornalístico. Para o gênero acadêmico, utilizou a Selva Científica (também parte da Floresta Sintá(c)tica). A avaliação apresentou acurácias de $98,06\%$, $98,30\%$ e $98,07\%$, respectivamente. Outros trabalhos baseados em RNRs e com uso de diferentes representações vetoriais alcançaram alto desempenho no corpus Bosque. Destacam-se o UDPipe 2 [Straka 2018], com $96,37\%$ de acurácia, o CNCSR [Heinzerling e Strube 2019], com $98,1\%$, e o Stanza [Qi et al. 2020], com $97,04\%$. Por fim, destaca-se o trabalho de [Bohnet et al. 2018], que utiliza a técnica de Meta-BILSTM, com a premissa de que o uso de diferentes representações vetoriais pode contribuir para o desempenho na tarefa. O modelo alcançou $98,11\%$ de acurácia no corpus Bosque.

Os conjuntos de etiquetas morfossintáticas (*tagsets*) variam nos diferentes trabalhos. Os trabalhos mais recentes fazem uso do *tagset* do modelo *Universal Dependencies* (UD) [de Marneffe et al. 2021], composto por 17 etiquetas. As classes abertas são representadas pelas etiquetas ADJ, ADV, INTJ, NOUN, PROPN e VERB; as classes fechadas são ADP, AUX, CCONJ, DET, NUM, PART, PRON e SCONJ; há também as etiquetas para outros casos, como PUNCT, SYM e X. O modelo UD já é adotado por mais de 100 línguas, contando com aproximadamente 200 *treebanks* catalogados. Esse modelo tem tido grande aceitação em função de sua proposta de "universalidade", com aplicação para línguas tipologicamente diferentes, já tendo passado por algumas versões. Como comentado anteriormente, este trabalho também se filia ao modelo UD.

3. Corpora

Neste trabalho, foram utilizados três corpora de gêneros diferentes, anotados manualmente segundo o modelo UD. Para o gênero jornalístico, foi utilizada a porção "base" do *treebank* Porttinari [Pardo et al. 2021], com notícias do jornal Folha de São Paulo. A porção "base" é a semente com base na qual o restante do *treebank* foi anotado. Para o gênero de UGC, adotou-se o corpus DANTEStocks [Di Felippo et al. 2021], que contém *tweets* do mercado financeiro. Contemplando o gênero acadêmico, o corpus PetroGold [Souza et al. 2021] apresenta uma coletânea de textos da área de óleo e gás, provenientes de teses, dissertações e monografias. Na Tabela 1, para evidenciar os desafios da tarefa, é possível visualizar um exemplo manualmente anotado de sentença ou *tweet* de cada corpus (a etiqueta morfossintática é separada dos *tokens* pela barra).

A Tabela 2 mostra o total de sentenças e *tokens* de cada corpus. É possível observar que o corpus DANTEStocks tem uma quantidade menor de *tokens* quando comparado aos corpora Porttinari-base e PetroGold. Ressalta-se que os corpora DANTEStocks e Porttinari-base originalmente não possuem a divisão em conjuntos de treino, validação e teste. Dessa forma, para fins de avaliação e comparação justa entre métodos, foi realizada essa divisão com a amostragem aleatória, utilizando a proporção de 10% para validação e 20% para o conjunto de teste, resultando nos números mostrados na tabela.

Tabela 2. Estatísticas dos corpora utilizados

Corpus	Gênero	Treino	Validação	Teste	Sentenças	*tokens*
Porttinari-base	Jornalístico	5.894	585	1.668	8.420	168.400
DANTEStocks	UGC	2.833	413	802	4.048	81.048
PetroGold	Acadêmico	8.054	447	445	8.946	250.905

É interessante notar dois pontos adicionais sobre os corpora selecionados. Em primeiro lugar, eles contêm textos bastante diferentes entre si, tanto em gênero quanto domínio. Isso é importante para o teste que este artigo se propõe a fazer, de avaliar a capacidade multigênero dos métodos. Em segundo lugar, há outros corpora que são anotados com UD e disponibilizados publicamente, como o Bosque [Rademaker et al. 2017], o CINTIL [Branco et al. 2022] e o PUD (*Parallel Universal Dependencies*) [Zeman et al. 2017], mas que foram preteridos por não seguirem diretrizes de anotação similares e não conterem apenas textos em português brasileiro. Os três corpora selecionados, além de serem para o português brasileiro, fazem parte de um esforço nacional de estudo e uniformização de UD para o português[1]. Dessa forma, há menos variáveis envolvidas nos experimentos realizados.

4. Experimentos

A experimentação foi dividida em duas etapas. A primeira consistiu em avaliar diferentes técnicas de etiquetagem no corpus jornalístico, o Porttinari-base. Em seguida, aplicou-se no contexto multigênero a técnica de melhor desempenho, considerando então os demais corpora. Essa estratégia visou a otimizar a sequência de testes necessários.

4.1. Técnicas de etiquetagem morfossintática

Foram selecionadas sete técnicas/modelos de etiquetagem morfossintática para a avaliação no corpus Porttinari-base, sendo esta seleção feita com base na representatividade e no desempenho dessas técnicas na literatura.

O primeiro modelo, UDPipe 2 [Straka 2018], foi avaliado com o tamanho de lotes (*batch size*) de 128 amostras, com um treinamento de 16 épocas, onde, nas primeiras 8 épocas, é utilizada a taxa de aprendizagem de 10^{-3}, e de 10^{-4} nas demais. Como modelo de língua, foi utilizado o BERTimbau [Souza et al. 2020].

O Stanza [Qi et al. 2020] possui um módulo de etiquetagem morfossintática que utiliza redes Bi-LSTM para a classificação. Para este modelo, foi utilizado o tamanho em lotes padrão de 5.000, taxa de aprendizagem de 10^{-3} e número máximo de atualizações de etapas de gradiente de 1.000.

[1] `https://sites.google.com/icmc.usp.br/poetisa`

O terceiro modelo, Meta-BiLSTM [Bohnet et al. 2018], foi treinado com o tamanho de lotes de 40.000 para o modelo em nível de palavras e 80.000 para o modelo em nível de caracteres. A taxa de aprendizagem é de 2×10^{-3}, com 3 camadas ocultas com 400 neurônios cada. O modelo utiliza representações estáticas em nível de palavra, obtidas do Skip-gram do Word2Vec com dimensão 300 [Hartmann et al. 2017].

Outra técnica foi a CNCSR [Heinzerling e Strube 2019], que se baseia no uso de representações vetoriais em nível de palavra e caractere com rede Bi-LSTM. Foram utilizadas as representações em nível de caractere e subpalavra, sendo elas combinadas por meio de uma rede RNR meta. O modelo foi treinado com tamanho de lotes de 64, número de épocas mínimo de 50 e máximo de 1.000, taxa de aprendizagem de 10^{-4}, tamanho de vocabulário de 100.000 e taxa de *dropout* de $0,2$. O modelo em nível de caractere possui representação vetorial de tamanho 50 e camada oculta com 256 neurônios; os modelos de subpalavra e meta possuem o mesmo número de neurônios na camada oculta.

Além destes modelos, foram realizados experimentos com três diferentes modelos de língua em conjunto com etapas de ajuste fino. Dessa forma, são utilizadas as representações da primeira subpalavra de cada *token* da sentença de entrada para detectar a classe gramatical. Foram utilizados os modelos de língua BERTimbau [Souza et al. 2020], DeBERTa-v3 [He et al. 2021] e XLM-R [Conneau et al. 2020]. Para os três modelos, foram utilizados os seguintes hiper-parâmetros: máximo de 30 épocas, taxa de aprendizagem de 2×10^{-5} e *weight decay rate* de $0,01$, que é um parâmetro do otimizador AdamW [Loshchilov e Hutter 2019]. Os modelos BERTimbau e XLM-R utilizaram tamanho de lotes de 32 e, para o DeBERTa-v3, foi utilizado tamanho 16.

O procedimento experimental conta com a realização de 10 execuções[2] de treinamento no conjunto de treino do corpus Porttinari-base, para, então, realizar a comparação entre os modelos e realização de testes de hipótese para identificar diferenças estatisticamente significativas na acurácia. O teste Anova [Fisher 1992] com *post hoc* de Tukey [Tukey 1949] foi selecionado para realizar esta avaliação. O teste Anova avalia se existem diferenças significativas entre as médias de dois ou mais grupos. Se identificada tal diferença, o teste de Tukey é aplicado para determinar quais os grupos que possuem médias significativamente distintas entre si, com correção para múltiplas testagens.

4.2. Resultados

A Tabela 3 apresenta os resultados da avaliação da etiquetagem morfossintática no corpus jornalístico. São apresentadas a acurácia média e a Medida-F Macro média das 10 execuções de experimentos para cada abordagem avaliada, além dos respectivos desvios padrões. É possível observar que os métodos baseados em RNRs possuem desempenho inferior aos métodos baseados em modelos de língua com ajuste fino, tanto em termos de acurácia quanto em Medida-F macro. Além disso, a abordagem com o BERTimbau possui o maior valor absoluto médio para acurácia e Medida-F Macro. Os modelos DeBERTa-v3 e XLM-R possuem valores próximos. As diferenças observadas com relação à acurácia foram significativas (Anova $Z(70, 69) \approx 890, p = 6e - 59$), com nível de confiança de 95%). Em análise par-a-par, as diferenças observadas foram significativas para todos os pares, exceto para BERTimbau $\times$ DeBERTa-v3 e XLM-R $\times$ DeBERTa-v3.

[2]Cada experimento utilizou o mesmo conjunto de treinamento, com variação na semente aleatória que é utilizada na inicialização dos pesos do modelo.

Tabela 3. Acurácia no corpus jornalístico Porttinari-base

Modelo	Abordagem	Acurácia média (%)	Medida-F macro média (%)
BERTimbau	Modelo de língua	**99,07 ± 0,03**	**96,39 ± 0,32**
DeBERTa-v3	Modelo de língua	$99,02 \pm 0,05$	$95,81 \pm 0,39$
XLM-R	Modelo de língua	$99,00 \pm 0,04$	$96,36 \pm 0,42$
Meta-BiLSTM	RNR	$98,47 \pm 0,06$	$94,89 \pm 0,28$
Udpipe 2	RNR	$98,01 \pm 0,03$	$93,13 \pm 0,54$
Stanza	RNR	$98,22 \pm 0,05$	$94,60 \pm 0,27$
CNCSR	RNR	$98,10 \pm 0,07$	$94,04 \pm 0,30$

Dado que não foi observada diferença significativa entre os métodos baseados nos modelos BERTimbau e DeBERTa-v3, o método baseado no BERTimbau foi selecionado para a próxima etapa de experimentação, devido a seu menor número de parâmetros. O método foi avaliado nos três corpora de gêneros diferentes (jornalístico, acadêmico e UGC), em que o experimento é constituído pelo treinamento do modelo em cada cenário de combinação dos corpora, seguido de sua avaliação separada em cada corpus individual. A Tabela 4 exibe a acurácia média dos experimentos nos conjuntos de teste.

Tabela 4. Acurácia no contexto multigênero

	Acurácia média (%)		
Corpora de treinamento	Porttinari-base	DANTEStocks	PetroGold
Porttinari-base	**99,07 ± 0,03**	87,14 ± 0,60	96,46 ± 0,17
DANTEStocks	96,55 ± 0,23	**97,98 ± 0,08**	94,95 ± 0,20
PetroGold	96,99 ± 0,10	84,96 ± 0,46	**98,93 ± 0,06**
Porttinari-base + DANTEStocks	99,05 ± 0,04	97,91 ± 0,10	96,58 ± 0,16
Porttinari-base + PetroGold	98,95 ± 0,06	85,29 ± 0,34	98,85 ± 0,07
DANTEStocks + PetroGold	97,86 ± 0,06	97,99 ± 0,07	98,92 ± 0,05
Port.-base + DANTEStocks + PetroGold	**99,00 ± 0,05**	**97,92 ± 0,13**	**98,89 ± 0,06**

É possível observar que o cenário que obteve a maior acurácia média foi o cenário onde o modelo foi treinado apenas com dados do gênero alvo. Por exemplo, o melhor cenário para o corpus de gênero acadêmico foi o cenário em que o treinamento foi exclusivamente neste gênero. Contudo, estes modelos possuem acurácias mais baixas nos outros gêneros, por exemplo, o modelo treinado no corpus PetroGold com acurácia de $98,93\%$ no gênero acadêmico possui acurácia de $84,96\%$ no gênero UGC.

Também se pode notar maior discrepância entre os gêneros que seguem a norma culta da língua e o gênero UGC, que possui características linguísticas diferentes. Quando o cenário com o PetroGold é avaliado no gênero jornalístico, por exemplo, é possível observar uma acurácia de $96,99\%$ (ou seja, há uma diferença relativamente pequena em relação ao melhor resultado para esse gênero). Já no gênero UGC, observa-se uma diferença maior em relação ao melhor modelo treinado no corpus DANTEStocks.

Em relação ao treinamento multigênero, é possível observar que o modelo treinado em todos os gêneros (última linha da tabela) alcançou desempenho similar aos modelos treinados isoladamente, sendo que a diferença entre as médias possui valor máximo de $0,067$. Como esperado, essa diferença não foi estatisticamente significativa [3].

[3]Anova $Z(70,69) \approx 1107, p = 7e-62$). Tukey: Multigênero vs Porttinari-base $Z \approx 0,7e-4, p \approx$

Além da acurácia em nível de *tokens*, também foi calculada e acurácia em nível de sentença nos corpora, computando-se a porcentagem de sentenças que foram anotadas de forma completamente correta, obtendo-se os seguintes resultados médios nos corpora: Porttinari-base – $64,59\%$; DANTEStocks – $54,25\%$; PetroGold – $47,36\%$; Porttinari-base + DANTEStocks – $68,31\%$; Porttinari-base + PetroGold – $55,01\%$; DANTEStocks + PetroGold – $72,79\%$; Porttinari-base + DANTEStocks + PetroGold – $77,70\%$. Novamente, o cenário multigênero destaca-se. Aprofundando o estudo, na análise das sentenças com erros no cenário multigênero, é possível observar que: em 77% das sentenças, houve apenas 1 erro; em 18%, dois erros; em 4%, 3 ou 4 erros; o restante ($< 1\%$) tem 5 ou mais erros (que incluem casos de sentenças de estrutura incomum). Os resultados indicam um novo estado da arte para a língua portuguesa, além de demonstrarem que é possível ter um sistema multigênero robusto que possibilite o desenvolvimento de aplicações de PLN mais generalistas e que possam ser aplicados para textos variados.

Após a avaliação quantitativa, partiu-se para a avaliação qualitativa, essencial para entender a potencialidade real desse tipo de sistema e suas limitações. Partindo do modelo treinado no contexto multigênero, foi realizada a análise manual de erros (com o apoio de um linguista experiente), buscando-se encontrar erros ocorridos para cada etiqueta morfossintática. Aqui são reportados apenas os erros sistemáticas observados.

Com relação à etiqueta ADJ, no corpus Porttinari-base, foram encontrados 23 casos onde os *tokens* estavam na forma de particípio. Particípio é uma forma nominal do verbo e pode assumir as etiquetas ADJ, NOUN ou VERB, sendo um caso particularmente desafiador para a Linguística [Duran 2021]. Naturalmente, o mesmo tipo de erro é encontrado ao analisar os erros das etiquetas NOUN e VERB. Nos corpora DANTEStocks e PetroGold, foram encontradas 4 ocorrências em ambas as análises.

Para a etiqueta PROPN, é possível identificar casos em que o modelo classificou como NOUN, consistindo em outra dificuldade conhecida da área. No corpus Porttinari-base, foram 10 ocorrências; no DANTEStocks, 21; e 8 ocorrências no PetroGold. Em especial, no DANTEStocks, foi observado que alguns *tweets* continham índices da bolsa de valores sendo classificados com a etiqueta X. No total, foram encontradas 30 ocorrências desse tipo. Esse corpus adotou a etiqueta X para índices da bolsa que não possuíam função linguística no *tweet* e, quando possuíam função, a etiqueta PROPN deveria ser utilizada.

Finaliza-se com a etiqueta X, utilizada para casos a que outras etiquetas não podem ser associadas. No corpus Porttinari-base, todos os erros encontrados foram casos de estrangeirismos a que o modelo tentou associar uma classe gramatical diferente da etiqueta X. Este tipo de erro foi encontrado em 11 casos no corpus DANTEStocks e não ocorreu no corpus PetroGold. Esse é um erro considerado plausível, já que estrangeirismos poderiam ter outras etiquetas associados a eles.

É interessante observar que, caso esses erros relatados fossem computados como análises plausíveis no cálculo da acurácia, a acurácia geral do melhor modelo de etiquetagem se aproximaria dos 100%. Esses casos também podem servir de base para futuras discussões e eventuais aprimoramentos nos corpora anotados.

$0,77$, Multigênero vs DANTEStocks $Z \approx 0,7e{-}4, p \approx 0,99$, Multigênero vs PetroGold $Z \approx 0,4e{-}4, p \approx 0,99$ ao nível de confiança de 95%.

5. Experimentos adicionais: o corpus Mac-Morpho

Dada a relevância histórica do corpus Mac-Morpho [Aluísio et al. 2003] para a tarefa de etiquetagem morfossintática para o português, testamos nesse corpus a melhor técnica de etiquetagem observada no experimento anterior. O Mac-Morpho contém cerca de 1 milhão de palavras em português brasileiro, criado a partir de textos de jornais e revistas. A versão atual, Mac-Morpho v2 [Fonseca e Rosa 2013], conta com 23 etiquetas morfossintáticas de base e 7 complementares. Sendo assim, o conjunto de etiquetas é distinto do conjunto da UD. A contribuição desses experimentos adicionais reside, portanto, na avaliação da robustez da melhor técnica identificada em dados com um *tagset* diferente.

A Tabela 5 exibe as acurácias obtidas por trabalhos prévios da literatura e pelo etiquetador deste artigo baseado no modelo BERTimbau. É possível observar que o etiquetador deste trabalho obteve a maior acurácia, demonstrando sua robustez e avançando o estado da arte de etiquetagem para o corpus Mac-Morpho também.

Tabela 5. Acurácia para o corpus Mac-Morpho

Método	[Fonseca e Rosa 2013]	[de Sousa e Lopes 2019]	[Fonseca et al. 2015]	[Santos e Zadrozny 2014]	BERTimbau
Acurácia	96,48%	97,62%	97,31%	97,47%	**98,36%**

6. Considerações finais

Este trabalho avançou a fronteira do conhecimento e o estado da arte ao demonstrar a potencialidade multigênero de um método de etiquetagem morfossintática baseado em modelagem de língua e ao produzir resultados superiores ao estado da arte.

O melhor método observado, baseado no modelo BERTimbau, demonstrou uma boa capacidade de generalização nos gêneros abordados, mas pode ser interessante no futuro avaliá-lo ainda em outros gêneros e domínios a fim de confirmar tal robustez. Outro fator importante a ser considerado é o custo computacional desse etiquetador. Possuindo cerca de 110 milhões de parâmetros e complexidade quadrática no mecanismo de autoatenção, o tempo de inferência é considerável. Pode ser interessante explorar técnicas de compressão de modelos para reduzir o tamanho e tempo de inferência.

Para reprodução dos resultados apresentados, o repositório[4] de código é disponibilizado. Além disso, uma aplicação[5] foi criada para que interessados possam utilizar o melhor etiquetador desenvolvido (no cenário multigênero ou não). Outras informações sobre este trabalho e sobre iniciativas relacionadas podem ser encontradas no portal web do projeto POeTiSA[6].

Agradecimentos

Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI - `http://c4ai.inova.usp.br/`), com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM. Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei N. 8.248, de 23 de outubro de 1991, no âmbito do PPI-Softex, coordenado pela Softex e publicado como Residência em TIC 13, DOU 01245.010222/2022-44.

[4] `https://github.com/huberemanuel/porttagger`

[5] `https://huggingface.co/spaces/Emanuel/porttagger`

[6] `https://sites.google.com/icmc.usp.br/poetisa/`

Referências

Afonso, S., Bick, E., Haber, R., e Santos, D. (2002). Floresta sintá(c)tica: A treebank for Portuguese. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 1698–1703, Las Palmas, Spain.

Aluísio, S., Pelizzoni, J., Marchi, A. R., de Oliveira, L., Manenti, R., e Marquiafável, V. (2003). An account of the challenge of tagging a reference corpus for brazilian portuguese. In *6th international conference on Computational processing of the Portuguese language*, page 110–117, Faro, Portugal.

Bick, E. (2000). *The Parsing System "Palavras". Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. University of Arhus.

Bohnet, B., McDonald, R., Simões, G., Andor, D., Pitler, E., e Maynez, J. (2018). Morphosyntactic tagging with a meta-BiLSTM model over context sensitive token encodings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2642–2652, Melbourne, Australia.

Branco, A. e Silva, J. (2004). Evaluating solutions for the rapid development of state-of-the-art POS taggers for Portuguese. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, pages 507–510, Lisbon, Portugal.

Branco, A., Silva, J. R., Gomes, L., e António Rodrigues, J. (2022). Universal grammatical dependencies for Portuguese with CINTIL data, LX processing and CLARIN support. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5617–5626, Marseille, France.

Cabral, B., Souza, M., e Claro, D. B. (2022). Portnoie: A neural framework for open information extraction for the portuguese language. In *Computational Processing of the Portuguese Language: 15th International Conference*, page 243–255, Berlin, Heidelberg.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., e Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online.

de Marneffe, M.-C., Manning, C. D., Nivre, J., e Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47:255–308.

de Sousa, R. C. C. e Lopes, H. (2019). Portuguese pos tagging using blstm without handcrafted features. In Nyström, I., Hernández Heredia, Y., e Milián Núñez, V., editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 120–130, Havana, Cuba.

Di Felippo, A., Postali, C., Ceregatto, G., Gazana, L., Silva, E., Roman, N., e Pardo, T. (2021). Descrição preliminar do corpus dantestocks: Diretrizes de segmentação para anotação segundo universal dependencies. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 335–343, Porto Alegre, RS, Brasil.

Domingues, M. L. C. S. (2011). *Abordagem para o desenvolvimento de um etiquetador de alta acurácia para o Português do Brasil*. PhD thesis, Universidade Federal do Pará, Belém, PA, Brasil.

Duran, M. S. (2021). Manual de anotação de PoS tags: Orientações para anotação de etiquetas morfossintáticas em língua portuguesa, seguindo as diretrizes da abordagem universal dependencies (UD). Technical report, Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo, São Carlos, Brasil.

Fisher, R. A. (1992). *Statistical Methods for Research Workers*. Springer New York.

Fonseca, E. R., G Rosa, J. L., e Aluísio, S. M. (2015). Evaluating word embeddings and a revised corpus for part-of-speech tagging in portuguese. *Journal of the Brazilian Computer Society*, 21:1–7.

Fonseca, E. R. e Rosa, J. L. G. (2013). Mac-morpho revisited: Towards robust part-of-speech tagging. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, pages 1–10, Fortaleza, Brasil.

Garimella, A., Banea, C., Hovy, D., e Mihalcea, R. (2019). Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498, Florence, Italy.

Hartmann, N. S., Fonseca, E. R., Shulby, C. D., Treviso, M. V., Rodrigues, J. S., e Aluísio, S. M. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Anais do XI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 122–131, Porto Alegre, Brasil.

He, P., Gao, J., e Chen, W. (2021). Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543:1–19.

Heinzerling, B. e Strube, M. (2019). Sequence tagging with contextual and non-contextual subword representations: A multilingual evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 273–291, Florence, Italy.

Jurafsky, D. e Martin, J. H. (2009). *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall.

Lin, Y., Wang, C., Song, H., e Li, Y. (2021). Multi-head self-attention transformation networks for aspect-based sentiment analysis. *IEEE Access*, 9:8762–8770.

Lopes, L., Duran, M., Fernandes, P., e Pardo, T. (2022). PortiLexicon-UD: a Portuguese lexical resource according to Universal Dependencies model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6635–6643, Marseille, France.

Loshchilov, I. e Hutter, F. (2019). Decoupled weight decay regularization. In *7th International Conference on Learning Representations*, pages 1–19, Toulon, France.

Pardo, T., Duran, M., Lopes, L., Felippo, A. D., Roman, N., e Nunes, M. (2021). Porttinari - a large multi-genre treebank for brazilian portuguese. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 1–10, Porto Alegre, Brasil.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., e Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of*

the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 101–108, Online.

Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., e de Paiva, V. (2017). Universal Dependencies for Portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics*, pages 197–206, Pisa,Italy.

Santos, C. D. e Zadrozny, B. (2014). Learning character-level representations for part-of-speech tagging. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1818–1826, Bejing, China.

Souza, E., Silveira, A., Cavalcanti, T., Castro, M., e Freitas, C. (2021). Petrogold – corpus padrão ouro para o domínio do petróleo. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 29–38, Porto Alegre, Brasil.

Souza, F., Nogueira, R., e Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems*, pages 403–417, Cham.

Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, 5:99–114.

Zeman, D., Popel, M., Straka, M., Hajic, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gokirmak, M., Nedoluzhko, A., Cinkova, S., Hajic jr., J., Hlavacova, J., Kettnerová, V., Uresova, Z., Kanerva, J., Ojala, S., Missilä, A., Manning, C. D., Schuster, S., Reddy, S., Taji, D., Habash, N., Leung, H., de Marneffe, M.-C., Sanguinetti, M., Simi, M., Kanayama, H., dePaiva, V., Droganova, K., Martínez Alonso, H., Çöltekin, c., Sulubacak, U., Uszkoreit, H., Macketanz, V., Burchardt, A., Harris, K., Marheinecke, K., Rehm, G., Kayadelen, T., Attia, M., Elkahky, A., Yu, Z., Pitler, E., Lertpradit, S., Mandl, M., Kirchner, J., Alcalde, H. F., Strnadová, J., Banerjee, E., Manurung, R., Stella, A., Shimada, A., Kwak, S., Mendonca, G., Lando, T., Nitisaroj, R., e Li, J. (2017). Conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada.

Zhao, F., Quan, B., Yang, J., Chen, J., Zhang, Y., e Wang, X. (2019). Document summarization using word and part-of-speech based on attention mechanism. *Journal of Physics: Conference Series*, 1168:32008.